1. Circle which of the following are data mining tasks? (3 points)

a) Sorting a student database based on student identification numbers.

b) Monitoring the heart rate of a patient for abnormalities

c) Computing the total sales of a company

d) Store all data in an Excel file

e) Extracting the frequencies of a sound wave

f) Monitoring seismic waves for earthquake activities.

2. Classify the following attributes as binary, discrete, or continuous. Also classify them as nominal, ordinal, interval, or ratio. (5 points)

Example: Age in years. Answer: discrete, ratio.

4'

a) Bronze, Silver, and Gold medals as awarded at the Olympics.

discrete, ordinal

b) Number of patients in a hospital.

continuous, nominal

c) Military rank.

discrete, ordinal

d) Brightness as measured by a light meter.

continuous, ordinal

e) Angles as measured in degrees between 0^o and 360^o.

discrete, interval.

PLEASE KEEP YOUR EYES ON YOUR OWN PAPER

+7 3. For the following data objects, *i* and *j*, calculate the indicated similarity or distance measures. (8 points)

$i = (0, 1, 0, 1)$, $j = (1, 0, 1, 0)$.

2 Cosine similarity:

$$\text{Cosine similarity}(i,j) = \frac{i \cdot j}{\|i\| \times \|j\|}$$

$$= \frac{(0\times1 + 1\times0 + 0\times1 + 1\times0)}{1\times1} = \frac{0}{1} = 0$$

3 Eucledian distance:

$$d = \sqrt{\Sigma(x_{i_i} - x_{si})^2}$$

$$= \sqrt{(0-1)^2 + (1-0)^2 + (0-1)^2 + (1-0)^2}$$

$$= \sqrt{4}$$

$$= 2$$

2 Correlation (Pearson's) coefficient:

$$\text{correlation} = \frac{\sigma_{i,j}}{\sigma_i \sigma_j} \quad \text{(covariance of } i,j\text{)} \quad \text{(variance of } i,j\text{)}$$

$$\sigma_i = \frac{1}{N}\Sigma (x_i - \mu)^2 = \frac{-0.5 + 0.5 + -0.5 + 0.5}{4} = 0.$$

correlation = covariance = $E(i,j) - E(i)\,E(j)$ correlation = $\frac{0.75}{0}$

$= 1 - (0.5 \times 0.5)$ correlation = 0

$= 1 - 0.25$

$= 0.75$

PLEASE KEEP YOUR EYES ON YOUR OWN PAPER

10

4. Naïve Bayes Classifier (10 points)

Consider the following data set with Attributes A, B, C and class label "-" and "+" .

Index	A	B	C	Class
1	0	0	1	–
2	1	0	1	+
3	0	1	0	–
4	1	0	0	–
5	1	0	1	+
6	0	0	1	+
7	1	1	0	–
8	0	0	0	–
9	0	1	0	+
10	1	1	1	+

(a) Predict the class label for a test sample (A = 1, B = 1, C = 1) using the naive Bayes approach (8 points)

$P(C|X) = P(C) \prod P(X_i|C)$

$P(-) = 5/10 = 0.5$

$P(+) = 0.5$

$P(A|+) = 3/5$

$P(A|-) = 2/5$

$P(B|+) = 2/5$

$P(B|-) = 2/5$

$P(C|+) = 4/5$

$P(C|-) = 1/5$

$P(+|X) = 0.5\left(\frac{3}{5} \times \frac{2}{5} \times \frac{4}{5}\right)$

$= \frac{12}{5^3}$

$P(-|X) = 0.5\left(\frac{2}{5} \times \frac{2}{5} \times \frac{1}{5}\right)$

$= \frac{2}{5^3}$

∴ Class label will be +

(b) What is an assumption when using the Naïve Bayes classifier? (2 points)

The classifier assumes that the attributes are conditionally independent of each other.

PLEASE KEEP YOUR EYES ON YOUR OWN PAPER

5. Classification using Decision Trees (10 points)

Given the following training data, generate a decision tree that is at least three levels. Use the Gini index as the splitting criteria. Show the steps.

F1	F2	F3	CLASS
1	1	1	A
1	1	0	B
0	1	0	A
1	0	1	B

Gini Index (D) = $1 - \sum p_i$

Gini (CLASS) = $1 - \left[\frac{2}{4} + \frac{2}{4}\right] = 1 - [0.5 + 0.5] = 0$

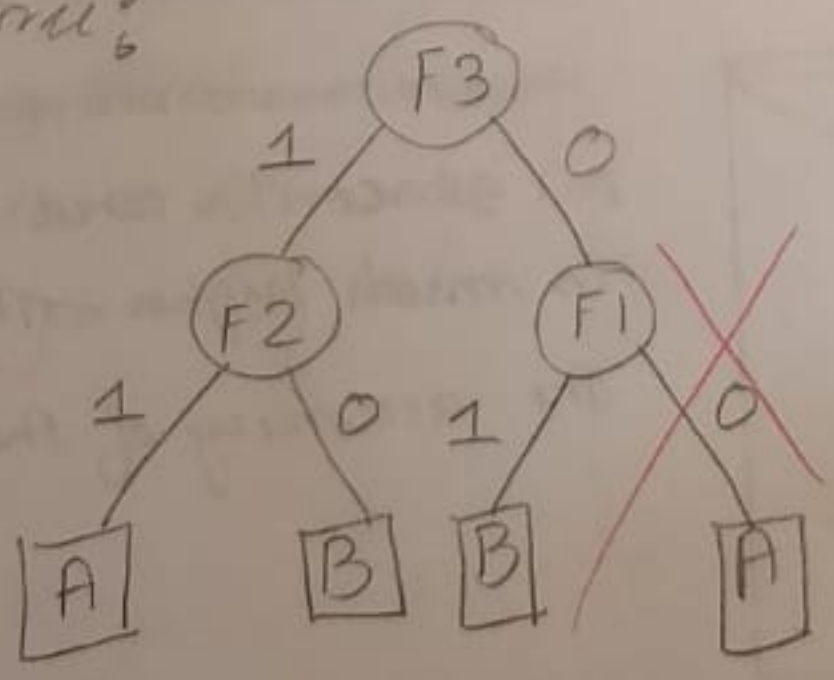
Gini (F1) = $\frac{3}{4}$ Gini (F1=1) + $\frac{1}{4}$ Gini (F1=0)

$= \frac{3}{4}\left\{1 - \left(\frac{1}{3} + \frac{2}{3}\right)\right\} + \frac{1}{4}\left\{1 - \left(\frac{1}{1} + \frac{0}{1}\right)\right\}$

Gini (F2) = $\frac{3}{4}$ Gini (F2=1) + $\frac{1}{4}$ Gini (F2=0)

Gini (F3) = $\frac{2}{4}$ Gini (F3=1) + $\frac{2}{4}$ Gini (F3=0)

Generated tree:



PLEASE KEEP YOUR EYES ON YOUR OWN PAPER

+10 6. Classifier performance and ROC curve (10 points)

For the following confusion matrix, calculate the following performance metrics

Prediction/ Classification	Truth: A	B	C	
A	90	9	1	100
B	6	86	8	100
C	4	5	91	100

TP FN
FP TN

(a) Accuracy of the classifier:

Accuracy = $\frac{90+86+91}{300}$

(b) Precision for each of the class labels A, B, C

Precision (P) = $\frac{TP}{TP+FP}$

$P(A) = \frac{90}{90+10} = 0.9$

$P(B) = \frac{86}{86+6+8}$

$P(C) = \frac{91}{91+4+5}$

(c) Recall for each of the class labels A, B, C

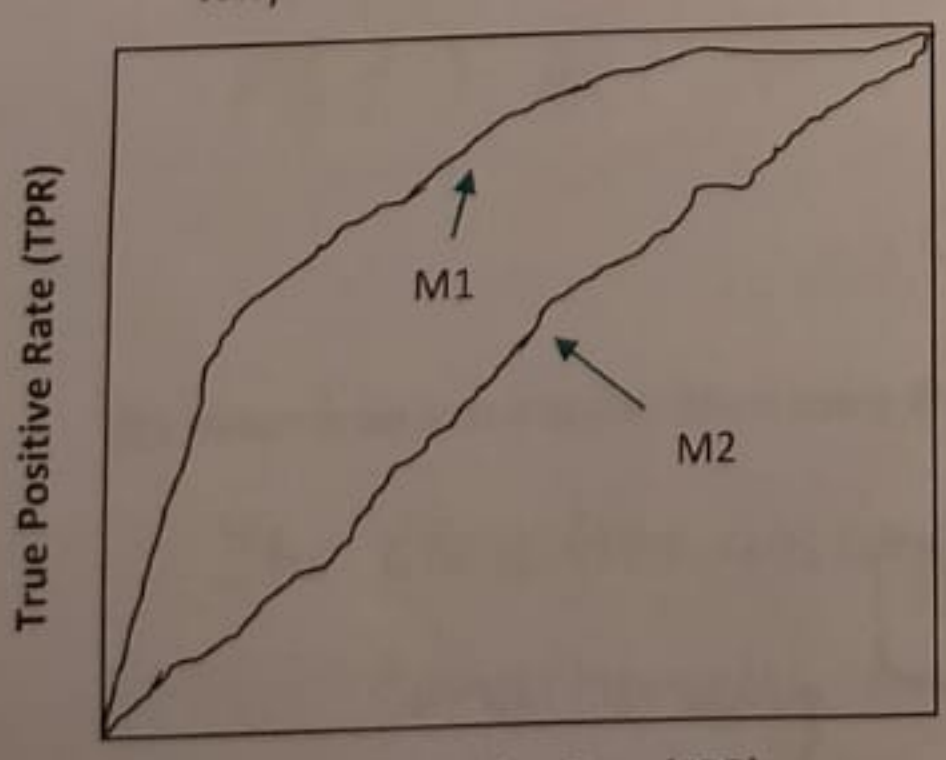
Recall (R) = $\frac{TP}{TP+FN}$

$R(A) = \frac{90}{90+9+1}$

$R(B) = \frac{86}{86+9+5}$

$R(C) = \frac{91}{91+8+1}$

(d) The ROC curves of two models (M1 and M2) are shown below. Which model is better? Why?



(Please use this space to enter your answer)

M1 since it's area under the curve is much higher. This area reflects the accuracy of the model.

PLEASE KEEP YOUR EYES ON YOUR OWN PAPER

7. Clustering Approaches (6 points)

6'

(a) List four clustering methods. Give one example of each method.

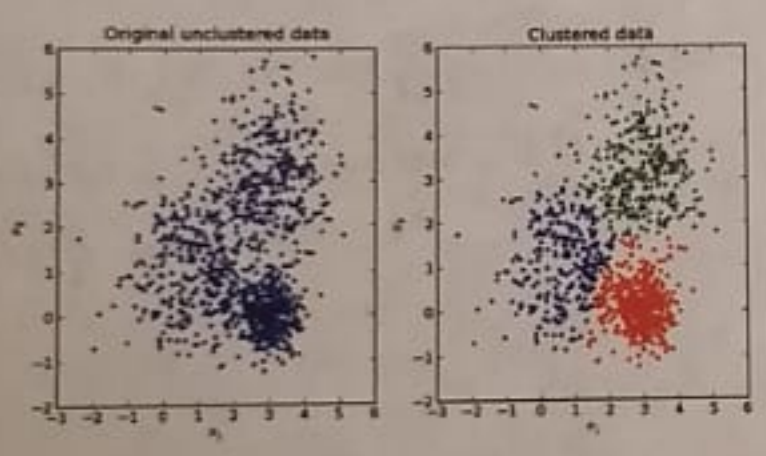
Partitioning methods - K-means,

Hirarchical methods - AGNES (Aglomeritive Nesting)

Density based methods - DBSCAN

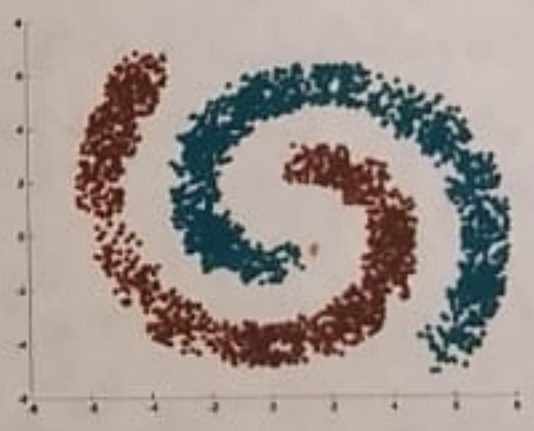
Grid based methods - STING

(b) For each of the data sets shown in (i) and (ii) below, describe which clustering method would be best suited? Why?



(i)

K-mechods since distance would be a good clustering measure and this method is more resiliant to noisy data.



(ii)

DBSCAN, since density would be a better clustering measure in this case.

PLEASE KEEP YOUR EYES ON YOUR OWN PAPER

8. Frequent Itemset Mining (8 points)

Customer	Transaction ID	Items Bought
1	1	{a, d, e}
1	24	{a, b, c, e}
2	12	{a, b, d, e}
2	31	{a, c, d, e}
3	15	{b, c, e}
3	22	{b, d, e}
4	29	{c, d}
4	40	{a, b, c}
5	33	{a, d, e}
5	38	{a, b, e}

7

1. Compute the support for itemsets {e}, {b, d}, and {b, d, e} for the transaction table provided above. (4 points)

Assuming we measure the absolute support not relative support. But relative support would be absolute support devided by number of itemsets (10).

$s(\{e\}) = 8$

$s(\{b,d\}) = 2$

$s(\{b,d,e\}) = 2$

-1

2. Use the results in part (1) to compute the confidence for the association rules {b, d} → {e} and {e} → {b, d}. (4 points)

confidence of $(\{b,d\} \rightarrow \{e\}) = \frac{2}{2} = 1$

confidence of $(\{e\} \rightarrow \{b,d\}) = \frac{2}{8} = 0.25$

PLEASE KEEP YOUR EYES ON YOUR OWN PAPER